時間周波数表現の標本化から音声の分析変換合成を考える (時間周波数解析の理論とその理工学的応用)

AUTHOR(S):

河原, 英紀

CITATION:

河原, 英紀. 時間周波数表現の標本化から音声の分析変換合成を考える (時間周波数解析の理論とその理工学的応用). 数理解析研究所講究録 2012, 1803: 12-25

ISSUE DATE:

2012-08

URL:

http://hdl.handle.net/2433/194372

数理解析研究所講究録
第1803巻 2012年 12-25

時間周波数表現の標本化から音声の分析変換合成を考える

Speech analysis, modification and synthesis in terms of sampling time-frequency representations

和歌山大学 河原 英紀
Hideki Kawahara, Wakayama University

E-mail:kawahara@sys.wakayama-u.ac.jp

1 はじめに

アナログ信号のディジタル化の基礎である標本化定理の広く知られた定式化[1]から、既に60年以上が経過した。この間に、ディジタル信号処理は広く普及するとともに、最近では、標本化定理をより自由な立場から拡張する動きが進んでいる[2, 3]。[1] 70年以上前のVoderおよびVocoder[5]に端を発する電気的音声処理技術も、この動きと呼応して、大きく変わろうとしている。ここでは、著者らが研究を進めている音声分析変換合成システムSTRAIGHT[6, 7]を中心として、標本化と音声分析合成の関係および今後の展開を考えてみたい。[2]

2 音声の時間周波数表現

70年前のVoder/Vocoderも、現在のSTRAIGHTも、音声信号の時間周波数表現を利用していることに変わりはない。「rst vocoder」という言葉で検索すると、女性のオペレータがVoderを『演奏』している動画を見つけることができる。オペレータの指が動いて時々刻々と時間周波数表現の形を変化させると、想像以上に明瞭な言葉が紡ぎ出される。オペレータは、指による「演奏」と同時に、足でペダルを操作して声の抑揚を加え、手首で有声音と摩擦音の音源の切換えを行なっている。これは、音声波形というアナログ信号中に早い変化（電話では毎秒約4000回）が含まれているとしても、話されている内容は、遥かに遅い変化（指の運動は、最速でも

[1]なお、これらの背景となる枠組みは、既に1980年代に小川によって提案されている。[4]

[2]STRAIGHTは、2008年にTANDEM-STRAIGHT[7]として、根本的に定式化し直されている。以下の議論では、TANDEM-STRAIGHTをSTRAIGHTと呼び、以前のものに言及する場合には、legacy-STRAIGHTと呼ぶこととする。

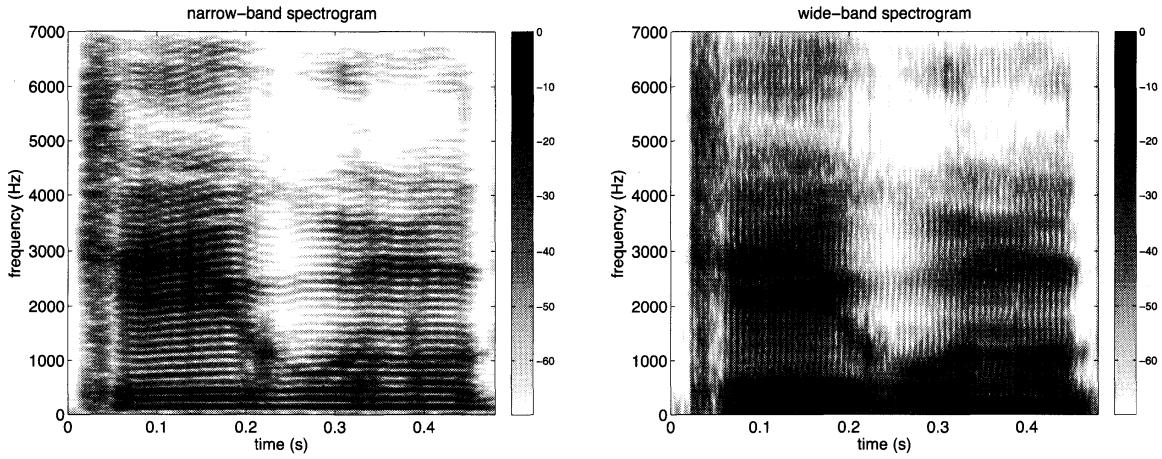

図 1: スペクトログラムの２つの表現。（左）狭帯域スペクトログラム、（右）広帯域スペクトログラム

毎秒数十回）で再現できることを意味する。音声の時間周波数表現に基づく、音声生成、知覚、分析、合成の研究の始まりである。

2.1 スペクトログラムに表れる特徴

これらで用いられている時間周波数表現（スペクトログラム $P(\omega,t)$）は、以下のように定義されている。

$$P(\omega,t) = \int_{\infty}^{\infty} w(\tau \quad t)x(\tau)e^{\ j\omega\tau}d\tau^{\ 2} \tag{1}$$

ここで、t は、時間、$x(t)$ は、分析対象となる信号、$w(t)$ は、信号を観測するための窓関数を表す。$w(t)$ には、通常は、$t=0$ に関して偶対称であり　$T_W/2 < t < T_W/2$ の区間で定義されている関数が用いられる。

スペクトログラムは、当初、帯域通過フィルタとヘテロダインを用いた中心周波数の走査を組み合わせたアナログ技術により計算されていた [8]。このような方法では、自由にフィルタの帯域幅を変更することができないため、図 1 に示すように、調波成分の分解を重視した狭帯域スペクトログラムと、時間的な変化の分解を重視した広帯域スペクトログラムが用いられていた。[3] 図では、男性が発話した「敬語の使い方は、難しいものです。」という文の最初の「敬語の」に相当する部分を示している。横軸を時間、縦軸を周波数とする表現が伝統的に用いられている。1950 年代までの研究により、文章音声の意味を伝えるのであれば、4000 Hz までの周波数成分を伝えれば十分であることが知られていた。[4] しかし、話者の個人性などの非言

[3] ディジタル信号処理の普及の原動力となった高速フーリエ変換は、当然ながら早速、スペクトログラムの計算に応用された [9]。

[4] 今でも電話の帯域は、300 Hz〜3,400 Hz と定められている。

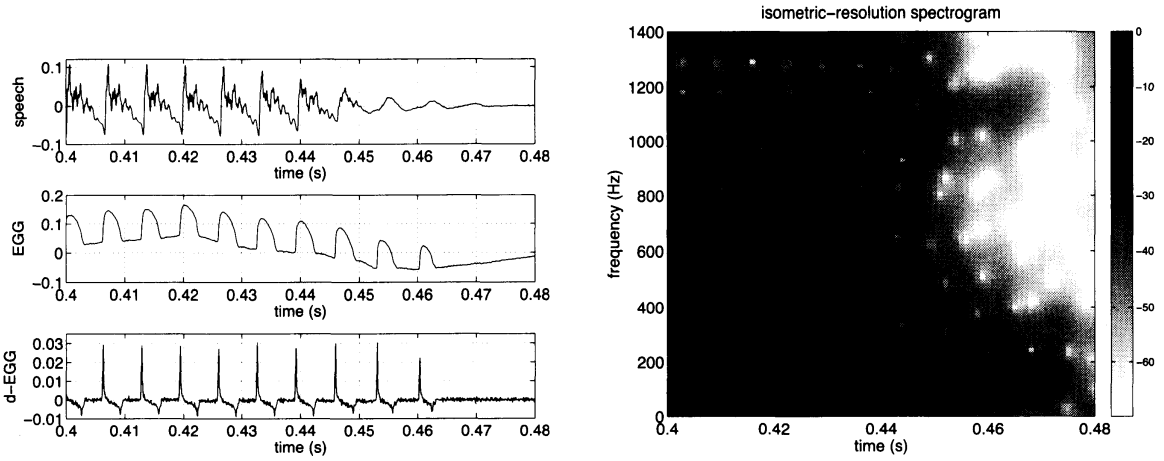

図 2: 音声波形とスペクトログラムの関係。(左図) 上から音声波形、EGG 信号、微分した EGG 信号、(右図) 同じ部分を、時間分解能と周波数分解能が、それぞれ基本周期と基本周波数に関して同程度になるように表示したスペクトログラム

語情報を十分に伝えるは、それ以上の成分も必要となる。[5]ここでは、7000 Hz までを表示している。

図 1 の左側の狭帯域スペクトログラムでは、基本波成分とその整数倍の周波数の成分に対応する横縞が認められ、右側の広帯域スペクトログラムでは、ほぼ周期的な縦縞が認められる。これらの構造に加え、ゆっくりとうねる太い横縞 (3 kHz までの範囲に 3 本程度) が重なっているように見える。これらの太い横縞は、信号のパワーが集中している領域を表しており、喉から唇に至る空洞 (声道) の形状により定まる共鳴の影響を反映している。これらの領域はホルマント (formant) と呼ばれ、低い周波数のものから順に第一、第二のように番号が付られている。ホルマントは、発声時の声道形状の違いを反映しており、異なった母音では異なった配置となる。子音の発声には、舌や顎や唇等の運動が伴うため、ホルマントの軌跡も時間とともに変化する。Voder のオペレータは、結局、指の操作によってこのような軌跡を作り出していたことになる。

2.2 標本化の手段としての有声音源

広帯域スペクトログラムに認められる縦縞は、有声音の生成の仕組みを反映している。有声音は、声門の開閉に伴う呼気流の断続により駆動される。図 2 の左図には、音声波形 (上段) と、声門の開閉を示す EGG 信号の波形 (中段)、および EGG(electroglottogram) 信号を時間について微分した信号 (下段) を示す。EGG 信号は、喉の左右に装着された電極の間を流れる電流の量を表している [10]。(EGG 信号の 0 という値自体には特別の意味はない。上の方では電流が多く流れ、下の方

[5]既存設備による歴史的な拘束を受けないインターネット上の音声通信の例では、8,000 Hz までの帯域が用いられている。このような広い帯域は、非母国語での通話を容易にする効果もある。

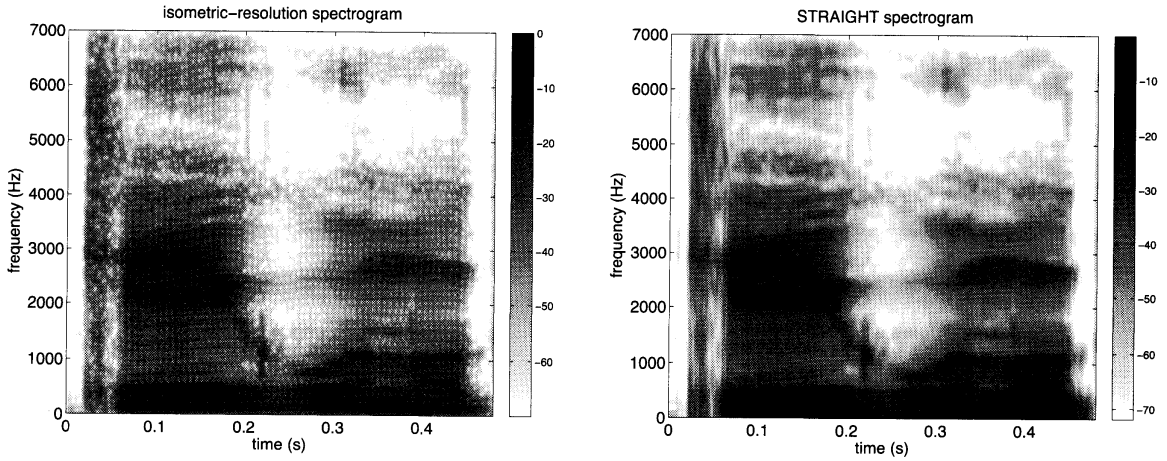

図 3: 周期的駆動の影響を含むスペクトログラム（左図）と、その影響を選択的に取り除いたスペクトログラム（右図：STRAIGHT スペクトログラム）

では電流が少ないことを示している。正負で電流の向きが反転する訳ではない。）左右の声帯が接触して声門が閉じているときには電流が多く流れ、声門が開いている場合には、声門を迂回する部分を流れる電流だけになる。したがって、図の山形の部分は、声門が閉じている区間を表していることになる。最下段の波形から、声門は徐々に開き、急に閉じてしまうことが分かる。

声門が急速に閉じてしまうことにより、呼気流が切断され、流速は突然 0 になる。この呼気流の不連続が、図 2 の左図の最下段の鋭いスパイクの時刻に生ずる。音声波形の高い周波数成分のパワーの大部分は、この不連続により供給されている。見方を変えると、時々刻々と変化する声道形状の情報が、この不連続によって、周期的に採取（標本化）されていることになる。

図 2 の右図は、窓関数の時間方向の広がり σ_t と周波数方向の広がり σ_ω が、それぞれ音声の基本周期 T_0 と基本周波数 $f_0 = 1/T_0$ に対して同じ比となるようにして求めたスペクトログラムを示す。なお、σ_t と σ_ω は、次式により定義される。

$$\sigma_t^2 = \frac{\int_\infty^\infty t^2 |w(t)|^2 dt}{\int_\infty^\infty |w(t)|^2 dt}, \quad \sigma_\omega^2 = \frac{\int_\infty^\infty \omega^2 |W(\omega)|^2 d\omega}{\int_\infty^\infty |W(\omega)|^2 d\omega}, \tag{2}$$

ここで、$W(\omega)$ は、$w(t)$ の Fourier 変換を表す。図 2 の右図の格子模様の縦の格子は、左図の最下段の鋭いスパイクの位置に対応していることが分かる。この格子模様が重なっている表現から、声道形状に対応する滑らかに変化する時間周波数表現を復元することが、解くべき問題である。要するに、図 3 の左側の図から右側の図を求めるのである。図 4 の音声生成過程に当てはめると、一番右側で観測された音声波形から、標本化される前の滑らかなフィルタ特性を求めることが解くべき問題となる。

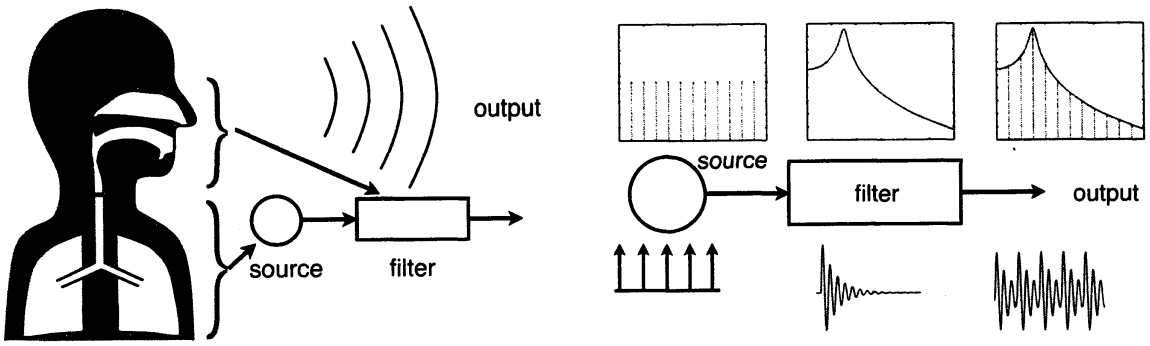

図 4: 音声生成過程。(左図) 肺と声帯から構成される音源部分と、その上部の空洞 (声道) で構成されるフィルタ部分により、音声は生成される。(右図) 有声音の周期的駆動は、フィルタの特性を周波数軸上で標本化する

3 滑らかな表現の復元

格子模様が重なっている表現からの復元は、2つのステップを通じて行なわれる。まず、分析位置に依存しないパワースペクトルが求められ、次いで、周波数領域での周期的構造が取り除かれる。数値的な細部にわたる議論は別の資料 [11] に譲り、ここでは標本化との関係を中心として議論を進める。

3.1 時間方向の周期的変動の除去

周期が T_0 である周期信号の場合、$T_0/2$ の間隔を隔てて求められたパワースペクトルを平均することで、パワースペクトルに含まれていた分析位置 t に依存する項が消去される [12]。こうして求められるパワースペクトル (TANDEM スペクトル) を $P_T(\omega, t)$ と表すことにする。

$$P_T(\omega, t) = \frac{P(\omega, t \quad \frac{T_0}{4}) + P(\omega, t + \frac{T_0}{4})}{2} \tag{3}$$

この方法は、様々な窓関数 [13] に応用することができる。窓関数の Fourier 変換 $W(\omega)$ が実質的に非零となる周波数帯域の幅が $2/T_0$ 以下であれば、この方法を用いることにより分析位置に依存する項が消去される。

3.2 周波数方向の周期的変動の除去

時間方向の標本化の影響が、TANDEM スペクトルを用いることにより排除されたため、残る問題は、周波数方向の周期的変化の除去だけになる。音声における有声音は、図 4 に示したように、声道のインパルス応答と周期的に配置された δ 関数

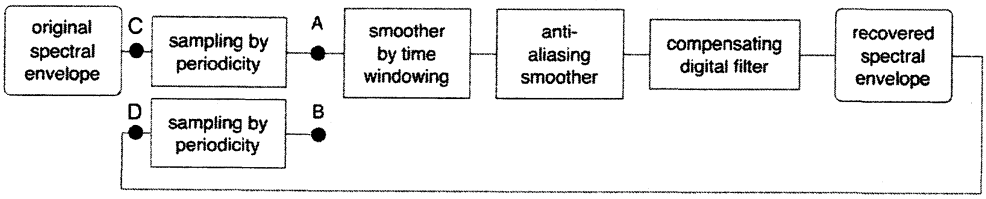

図 5: 周期信号による周波数標本化とスペクトル包絡の復元

の畳込みにより近似することができる。[6]このように近似すると、有声音のパワースペクトルは、声道のインパルス応答の Fourier 変換に相当する声道伝達関数を周波数軸上で基本周波数の整数倍の位置で標本化し、さらに、分析に用いた窓関数で平滑化したものの絶対値の自乗となることが分かる。[7]

この状況は、離散的な標本値から連続的な元のアナログ信号を復元する問題と同じである。TANDEM スペクトルでは、DA 変換後の低域通過フィルタが不完全なために、標本化に用いたパルスが残留していると考えれば良い。しかも、標本化の前に通常の AD 変換で用いられるアンチエリアシングフィルタを用いることができず、そのうえ、元の信号である声道伝達特性は、(空間周波数の意味で) 帯域制限されてはいない。良く知られた標本化定理を、この状況に用いることは適切ではない。

図 5 に、この状況を模式図で示す。図の C 点は、直接観測することができない滑らかな特性であり、有声音源による周期的駆動で基本周波数の整数倍の位置での値だけが A 点で得られることになる。実際に観測できるのは、窓関数の周波数領域での表現により平滑化されたものであり、これが、TANDEM スペクトルに相当する。解くべき問題は、この TANDEM スペクトルから元の観測できない滑らかな特性を復元することである。良く知られている標本化定理では、図の C 点と D 点の値が一致することを要請している。しかし、前述のように元の滑らかな特性は帯域制限されていないため、図の C 点と D 点の値を一致させることはできない。また、元の滑らかな特性を帯域制限して図の C 点の値とすることで D 点の値と一致させることは、音声の場合には帯域制限したスペクトルから復元した音声の品質劣化が大きいため、許されることではない。

3.2.1 consistent sampling

consistent sampling は、このような理想的ではない状況での標本化を対象としている。consistent sampling では、図の A 点と B 点における標本化された離散的な位置の値だけでの一致を要請し、C 点と D 点での値が一致することを要請しない。このように、従来の標本化定理よりも緩い要請とすることで、現実的な工学的手段を

[6]声帯音源波形の形状の影響や、声帯の振動と声道から反射して来る音波との非線形な相互作用は、無視する。

[7]TANDEM スペクトルの場合には、変動項が消去されるため、伝達関数の絶対値の自乗を、窓関数の Fourier 変換の絶対値の自乗で平滑化したものとなる。

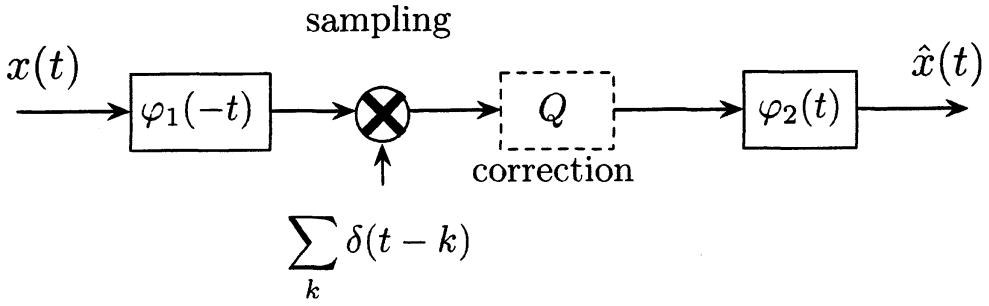

図 6: consistent sampling が対象とする信号の標本化と復元の構成

用いて、簡単に要請を満たすことが可能となる。図 5 に示す compensating digital lter が、そのための手段である。以下、文献 [14, 2] に基づいて説明する。

現実の機材を用いて標本化する場合を想定する。図 6 は、そのような機材で対象となるアナログ信号 $x(t)$ を標本化し、離散化された信号から再びアナログ信号 $\hat{x}(t)$ を復元する過程を表している。ここでは、信号は、まず標本化の前に機材の特性や前処理のための低域通過フィルタによる応答 φ_1 の影響を受ける。この影響を受けた信号が標本化され、離散信号となる。この離散信号に処理 Q が加えられ、得られた離散信号は、アナログ信号に変換するための機材の特性 φ_2 の影響を受けて、復元された信号 $\hat{x}(t)$ となる。φ_1 および φ_2 は、ここでは線形系のインパルス応答を想定している。文献 [14] の定理 1 は、文献 [2] では、以下の形で紹介されている。そこでは、信号 $x(t)$ を Hilbert 空間の要素 $x \in H$ として議論を進めている。$\langle\ \rangle$ は、L_2-ノルムを表す。以下、一部を省略して引用する。

> 応答 φ_1 の影響を受けた信号を標本化することは、φ_1 を用いて一連の内積 c_1 を求める（測定する）ことと同じである。
>
> $$c_1(k) = \langle x, \varphi_1(t \quad k) \rangle \tag{4}$$
>
> ここで、現実的な条件を考慮する（例えば $\varphi_1 \in L_2$）ことにより、$H = L_2$ として議論することができる。すると問題は、式 (4) という測定が行なわれたときに、次式で表される φ_2 による近似空間 $V(\varphi_2)$ により適切な近似を構成する問題になる。
>
> $$V(\varphi_2) = \left\{ s(t) = \sum_{k \in Z} c(k) \varphi_2(t \quad k) : c \in l_2 \right\} \tag{5}$$
>
> これは、結局、図 6 の Q として表されている適切な離散補償フィルタによる処理として解を与えることができる。
>
> (中略)
>
> 定理を説明する前に、相互相関の系列を次式により定義しておく。
>
> $$a_{12}(k) = \langle \varphi_1(t \quad k), \varphi_2(t) \rangle \tag{6}$$

定理：$x \in H$ を未知の入力された関数としたとき、ある m があって $|A_{12}(e^{j\omega})| \geq m$ となるとき、以下の意味で一貫性のある x の近似が $V(\varphi_2)$ の要素として一意に決まる。

$$\forall x \in H,\ c_1(k) = \langle x, \varphi_1(t \quad k)\rangle = \langle \tilde{x}, \varphi_1(t \quad k)\rangle \tag{7}$$

ここで信号 x の近似 $\hat{x}$ は、以下により与えられる。

$$\hat{x} = \hat{P}x(t) = \sum_{k\in Z}(c_1 \quad q)(k)\varphi_2(t \quad k)\ , \tag{8}$$

ここで、

$$Q(z) = \frac{1}{\sum_{k\in Z} a_{12}(k)z^{\ k}}\ , \tag{9}$$

であり、作用素 $\tilde{P}$ は、L_2 から $V(\varphi_2)$ への射影である。□

なお、

$$Q(z) = \sum_{k\in Z} q(k)z^{\ k}\ , \tag{10}$$

であることを補足しておく。

3.2.2 スペクトル包絡復元への応用

この consistent sampling をスペクトル包絡の復元に応用する場合には、φ_1 は、音声が厳密には周期的ではないことによるスペクトルの広がりに対応する。また、φ_2 は、時間窓を Fourier 変換して求められるスペクトルの平滑化関数と、調波構造を除去するための anti-aliasing フィルタの（周波数軸上での）応答関数の畳込に対応する。この anti-aliasing のための関数として、legacy-STRAIGHT では、2 次の cardinal-B spline の基底関数を用いており、TANDEM-STRAIGHT では、1 次の基底関数を用いている。

ただし、音声処理に応用する場合には、幾つかの近似と変換が必要となる。最初の条件は、復元された関数の非負性の保証である。TANDEM スペクトルは、パワースペクトルであり非負の値をとる。同様に、復元後の関数もパワースペクトルとして扱うためには、非負であることが必要となる。補償フィルタ Q の係数には負の値が含まれているため、音声のように成分の強度さが極端に大きい場合には、補償の結果に負の値が含まれる可能性がある。このような状況で、復元された関数の非負性を保証するために、以下の変換を利用する。まず、パワースペクトルを対数パワースペクトルに変換し、補償のための演算である Q をその上で行う。次に、こうして得られた結果を、指数関数を用いてパワースペクトルに戻す。この処理により、最初のパワースペクトルが正値であれば、復元された関数は正値となり非負性は保証される。[8]

[8]実際に観測される音声信号では、背景雑音などの存在によりパワースペクトルは、高い確率で正の値を取る。TANDEM スペクトルでは、正の値を取る確率はさらに高くなる。

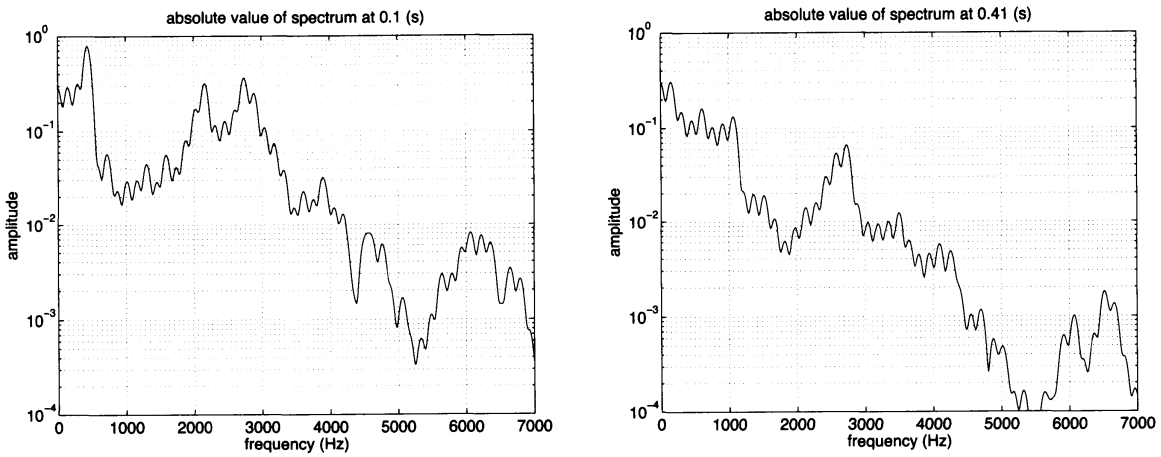

図 7: 実際の音声の TANDEM スペクトルの例。(左図)「敬語」の最初の/ke/の母音部/e/、(右図)「敬語の」の最後の/no/の母音部/o/

なお、この処理では、Q の係数をそのまま用いることができる。パワースペクトルの上に重畳している周期性に基づく変動が、その位置でのパワースペクトルの値に対して十分に小さい場合には、$\log(1+x) \approx x$ が良い近似となることを利用できるからである。状況を具体的に説明するために、図 7 に、実際に求められたパワースペクトルの平方根である絶対値スペクトルを示す。音声の基本周波数は、/e/の部分で 144.8 Hz, /o/の部分で 151.9 Hz である。スペクトル上の細かな凹凸は、この周期性によるものであり、それぞれの位置での値の ±20%程度あるいはそれ以下の大きさの変動として重畳していることが分かる。この重複分の大きさであれば、実用上、上記の近似を利用することに問題は無い。

実際に音声処理プログラムを作成する際には、無限個の係数がある Q をそのまま用いることはできない。幸いなことに、TANDEM スペクトルを求める際に用いる窓と anti-aliasing 用に用いる窓から求められる Q の係数の絶対値は、次数 k とともに急速に零に近づく。したがって、プログラムでは $k=0$ と $k=\pm 1$ に対応する係数のみを用いれば良い。結局、プログラムでは、以下の式により、TANDEM スペクトル $P_T(\omega,t)$ から、復元されたスペクトル $P_{TST}(\omega,t)$ (以下では、STRAIGHT スペクトル) を求めている。

$$P_{TST}(\omega) = \exp\ \mathcal{F}^{-1}[g_1(\tau)g_2(\tau)C_T(\tau)]\ , \tag{11}$$

$$\text{where}\quad g_1(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\ \frac{2\pi\tau}{T_0} \tag{12}$$

$$g_2(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau} = \mathcal{F}[h_2(\omega)], \tag{13}$$

$$h_2(\omega) = \begin{cases} 0 & |\omega| \quad \frac{\omega_0}{2} \\ \frac{1}{\omega_0} & \text{otherwise} \end{cases} \tag{14}$$

$$C_T(\tau) = \int_{\infty}^{\infty} \log\left(P_T(\omega,t)\right) e^{j\omega\tau} d\omega, \tag{15}$$

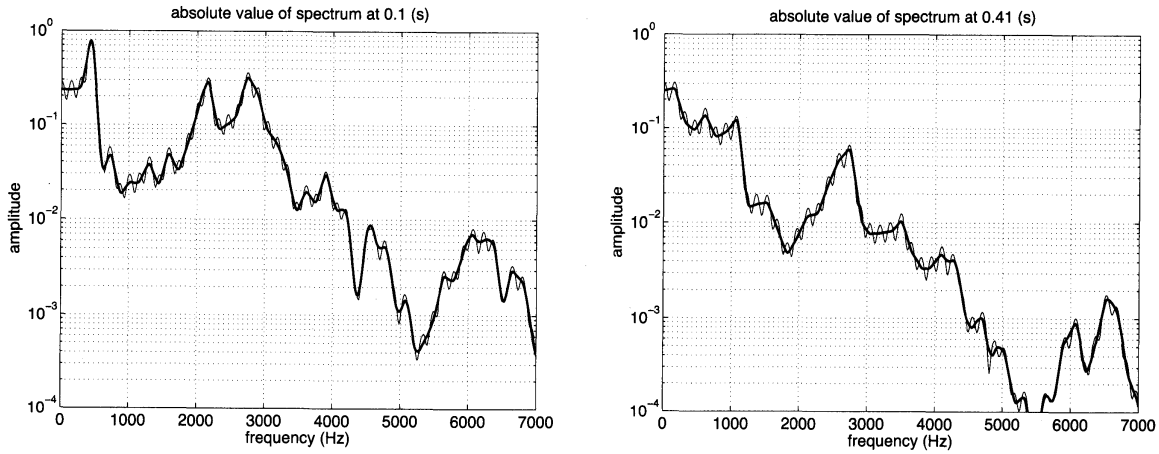

図 8: 実際の音声の TANDEM スペクトル（細線）と STRAIGHT スペクトル（太線）の例。（左図）「敬語」の最初の/ke/の母音部/e/、（右図）「敬語の」の最後の/no/の母音部/o/

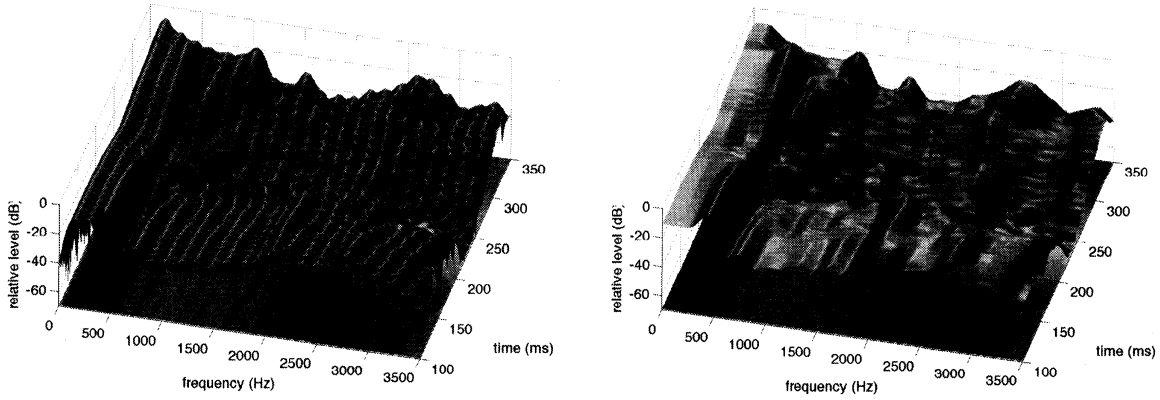

図 9: 実際の音声の短時間 Fourier 変換によるスペクトログラムの 3D 表示（左図）と STRAIGHT スペクトログラムの 3D 表示（右図）。

なお、ここで $\mathcal{F}$ は、Fourier 変換を表す。$\tilde{q}_0$ と $\tilde{q}_1$ は、式 (10) の q_0 と q_1 の値を、打切りの影響と最終的な合成音声への影響を考慮して調整した係数を表している [15]。

式 (15) で定義されるパワースペクトルの対数の Fourier 変換は、cepstrum（spectrum の綴り換えで「ケプストラム」と読む）と呼ばれる [16]。cepstrum は、時間としての意味を持つ τ の関数である。τ は、quefrency（これも同様に frequency の綴り換えで「ケフレンシ」と読む）と呼ばれる。$g_1(\tau)$ と $q_2(\tau)$ は、cepstrum への重み関数であり lifter(これも lter の綴り換え) と呼ばれる。

図 8 に、この方法によって図 7 に示した TANDEM スペクトルから求めた STRAIGHT スペクトルを示す。ここでは、STRAIGHT スペクトルと TANDEM スペクトルの自乗平均値が一致するように調整し表示している [17]。この図から分かるように、TANDEM スペクトルにあった基本周波数を周期とする周波数方向の変動は、STRAIGHT

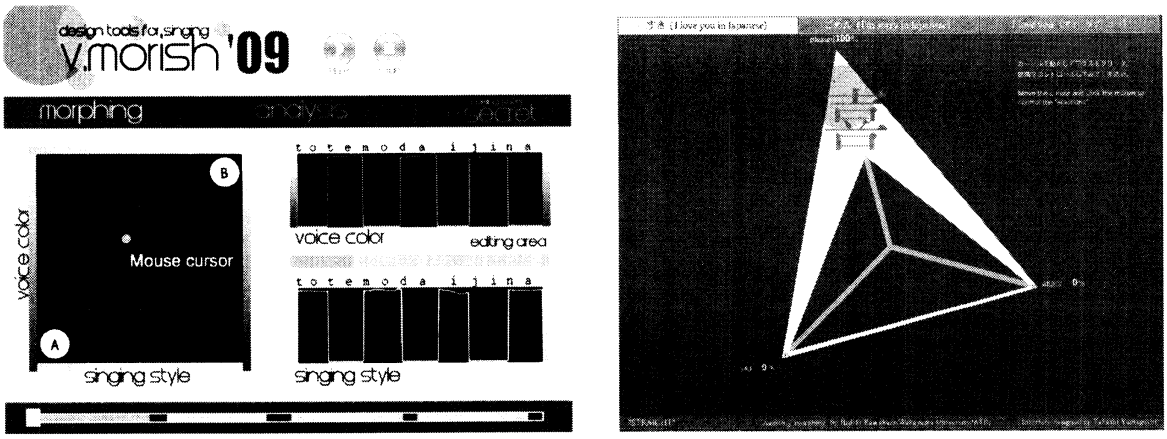

図 10: 音声モーフィングの応用例。(左図) vmorish'09 [20]、(右図) 日本科学未来館で展示されたデモ [21]。

スペクトルでは取り除かれている。図 9 に、同じ資料の 100 ms から 350 ms までの部分を拡大した 3D 表示で示す。STRAIGHT スペクトログラムでは、時間方向と周波数方向の双方の周期性の影響が取り除かれていることが分かる。

3.3 基本周波数の推定

ここで説明した方法では、基本周波数の値が利用される。しかし、基本周波数の値は通常は既知ではないため、まず音声信号から基本周波数を推定し、その推定値を利用して時間方向と周波数方向の変動を取り除くことになる。現在の STRAIGHT では、パワースペクトル上の周期的な変動から初期推定値を求め、瞬時周波数を用いて改良する方法が用いられている [11, 18]。ただし、この方法は、特殊な発声法を用いる歌唱音声や障害音声等、従来の方法では分析が困難な音声にも適用できる反面、多くの計算時間を要するという問題がある。この問題を解決するために、波形の対称性に基づく高速な計算方法を開発している [19]。

4 STRAIGHT の応用

こうして時間方向と周波数方向の双方の周期性の影響が取り除かれた STRAIGHT スペクトルを用いることで、これまでは困難だった様々な音声の処理が可能になる [22]。2 つの音声資料が与えられたときに、それぞれの資料の属性が様々な割合で混合 (内挿／外挿) された音声を合成する音声モーフィングは、その一例である [23, 24]。音声モーフィングは、もともとは人間が音声を知覚する仕組みを研究するため手段として開発されたが (応用例 [25, 26])、声優による様々な演技の表情を任意に混合したり、歌声の声質や歌い回しを加工するなど、新しいコンテンツを作るためのツールとして利用することもできる [20, 27]。

図 10 に v.morish'09 と日本科学未来館で展示されたモーフィングのデモの例を示す。v.morish'09 では、左の四角いパッドを用いて、A と B と記された二種類の歌唱の声質と歌い回しを、歌を再生しながらリアルタイムに独立に操作することができる。右側の上下のパネルは、それぞれの属性の操作量の時間変化を可視化したグラフである。日本科学未来館で展示されたデモは、「好き」「ごめん」「I love you.」を、声優が「喜」「怒」「哀」の三種類の感情を込めて演じた音声を素材にしている。画面に直接触れて指示することで、三角形が変形し、それらの感情を任意の割合で混合した音声が再生される。なお、このデモへのリンク [28] が公開されている。

周期性の影響が取り除かれた STRAIGHT スペクトルは、その他にも音声合成 [29] のためのデータ作成や、音声変換 [30] など様々な研究の基盤として用いられている。(GoogleSchlar [31] による STRAIGHT 関連研究の総引用数は 2012 年 4 月時点で 1,500 件程度である。) STRAIGHT の情報ページ [32] では、研究の最新の状況や様々な説明資料をリンクし紹介している。

5 まとめとこれから

ここでは、著者らが研究を進めている音声分析変換合成システム STRAIGHT を中心として、標本化と音声分析合成の関係について説明してきた。周期的駆動を、音声生成過程の背景にある滑らかな時間周波数表現を組織的に標本化する手段であると解釈することで STRAIGHT が発明された。これは、音声を表現する際のレベルを考慮すると、波形の一つ上のレベルでの標本化と考えることもできよう。表現のレベルは、ここが最後ではない。図 8 には、周期性による構造よりも大きな構造が重なっている様子が見えている。声道の共鳴の配置を表すホルマントによる構造である。さらに、時間方向には、言語という離散的な構造も加わる。これらの拘束を考慮すると、音声は、今回の周期性に基づく標本化よりももっと疎な標本化で十分に良く近似できる可能性があることが分かる。これらを、標本化の新しい観点である compressive sampling [33](あるいは compressed sensing [34]) として見直すことにより、これまでの音声科学 [35] の手法の、新しい意味が明らかになるかも知れない。柔軟な発想の若手の活躍に期待したい。

参考文献

[1] C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, Vol. 37, pp. 10–21, 1949.

[2] Michael Unser. Sampling–50 years after Shannon. *Proceedings of the IEEE*, Vol. 88, No. 4, pp. 569–587, 2000.

[3] Y. C. Elder and T. Michaeli. Beyond bandlimited sampling. *IEEE Signal Processing Magazine*, pp. 48–68, May 2009.

[4] H. Ogawa. A uni ed approach to generalized sampling theorems. *ICASSP1986*, pp. 1657–1660, 1986.

[5] Homer Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, Vol. 11, No. 2, pp. 169–177, 1939.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.

[7] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation. In *Proc. ICASSP 2008*, pp. 3933–3936. IEEE, 2008.

[8] W. Koenig, H. K. Dunn, and L. Y. Lacey. The sound spectrograph. *J. Acoust. Soc. Am.*, Vol. 18, pp. 19–49, 1946.

[9] Alan V. Oppenheim. Speech spectrograms using the fast fourier transform. *Spectrum, IEEE*, Vol. 7, No. 8, pp. 57 –62, aug. 1970.

[10] D. G. Childers, D. M. Hicks, G. P. Moore, and Y. A. Alsaka. A model for vocal fold vibratory motion, contact arean and the electroglottogram. *J. Acoust. Soc. Am.*, Vol. 80, No. 5, pp. 1309–1320, 1986.

[11] H. Kawahara and M. Morise. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modi cation and synthesis framework. *SADHANA - Academy Proceedings in Engineering Sciences*, Vol. 36, No. 5, pp. 713–722, 2011.

[12] 森勢将雅, 高橋徹, 河原英紀, 入野俊夫. 窓関数による分析時刻の影響を受けにくい周期信号のパワースペクトル推定法. 電子情報通信学会論文誌 D, Vol. J 90-D, No. 12, pp. 3265–3267, 2007.

[13] F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, Vol. 66, No. 1, pp. 51–83, 1978.

[14] M. Unser and A. Aldroubi. A general sampling theory for nonideal acquisition devices. *IEEE Trans. Signal Processing*, Vol. 42, No. 11, pp. 2915–2925, 1994.

[15] 阿竹義徳, 入野俊夫, 河原英紀, 陸金林, 中村哲, 鹿野清宏. 調波成分の瞬時周波数を用いた基本周波数推定方法. 電子情報通信学会論文誌 D, Vol. D-II-J83, No. 11, pp. 2077–2086, 2000.

[16] A. M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, Vol. 41, pp. 293–309, February 1967.

[17] 赤桐隼人, 森勢将雅, 入野俊夫, 河原英紀. スペクトルピークを強調した F0 適応型スペクトル包絡抽出法の最適化と評価. 電子情報通信学会 論文誌 A, Vol. J94-A, No. 8, pp. 557–567, 2011.

[18] 和田芳佳, 森勢将雅, 西村竜一, 入野俊夫, 河原英紀. 複数の周期成分を持つ音声のための周期構造抽出法と障害音声分析への応用について. 音響学会聴覚研究会資料, Vol. 41, No. 6, pp. 457–462, 2011.

[19] 河原英紀, 森勢将雅, 西村竜一, 入野俊夫. 基本波の FM と AM 成分に基づく高速な基本周波数推定法について. 音響学会聴覚研究会資料, Vol. 41, No. 9, pp. 679–684, 2011.

[20] M. Morise, M. Onishi, H. Kawahara, and H. Katayose. v.morish'09: A morphing-based singing design interface for vocal melodies. *Entertainment Computing–ICEC 2009*, pp. 185–190, 2009.

[21] 河原英紀（技術），山口崇（デザイン）．感情音声モーフィングデモ．日本科学未来館での展示, 4–8 2005.

[22] 河原英紀. Vocoder のもう一つの可能性を探る–音声分析変換合成システム straight の背景と展開–. 日本音響学会誌, Vol. 63, No. 8, pp. 442–449, 2007.

[23] Hideki Kawahara and Hisami Matsui. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Vol. I, pp. 256–259, Hong Kong, 2003.

[24] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In *Proc. ICASSP 2009*, pp. 3905–3908, 2009.

[25] Stefan R. Schweinberger, Christoph Casper, Nadine Hauthal, Juergen M. Kaufmann, Hideki Kawahara, Nadine Kloth, and David M.C. Robertson. Auditory adaptation in voice perception. *Current Biology*, Vol. 18, pp. 684–688, 2008.

[26] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G.A. Rousselet, H. Kawahara, and P. Belin. Vocal attractiveness increases by averaging. *Current Biology*, Vol. 20, No. 2, pp. 116–120, 2010.

[27] 河原英紀, 生駒太一, 森勢将雅, 高橋徹, 豊田健一, 片寄晴弘. モーフィングに基づく歌唱デザインインタフェースの提案と初期的検討. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3637–3648, 2007.

[28] http://www.wakayama-u.ac.jp/~kawahara/Miraikandemo/straightMorph.swf.

[29] H. Zen, K. Tokuda, and A.W. Black. Statistical parametric speech synthesis. *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[30] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[31] Google scholar: http://scholar.google.com/.

[32] http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_j.html.

[33] Emmanuel J. Candès. Compressive sampling. In *Proc. International Congress of Mathematicians*, Vol. 3, pp. 1434–1452, 2006.

[34] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, Vol. 52, No. 4, pp. 1289 –1306, april 2006.

[35] 藤村靖. 音声科学原論. 岩波書店, 2007.